INVESTIGATIVE RADIOLOGY
Volume 30, Number 6, 334–340

Guidelines for the Assessment of New Diagnostic Tests

YVONNE T. VAN DER SCHOUW, PHD,* ANDRÉ L. M. VERBEEK, MD, PHD,*
AND SJEF H. J. RUIJS, MD, PHD†

Schouw van der YT, Verbeek ALM, Ruijs SHJ. Guidelines for the assessment of new diagnostic tests. Invest Radiol 1995;30: 334–340.

RATIONALE AND OBJECTIVES. **Because new diagnostic tests become available rapidly, the authors determined a need for proper assessment of tests before their implementation in clinical practice. Three factors are of pivotal importance: the selection of the proper study population, the determination of the diagnostic power including its related statistical analysis, and the relation of the new test to current diagnostic tools. Patients suspected of having a disease are those who would benefit from the application of a new test. Therefore, only those patients need be involved in the assessment study.**

METHODS. **Summary measures of diagnostic power other than sensitivity and specificity are advocated because these conventional measures depend on cutoff points and are susceptible to selection bias. The relation between the new test and existing diagnostic tools must be established to determine if the new test contributes to the diagnostic process.**

RESULTS AND CONCLUSION. **To avoid waste of effort and money, the authors suggest a prudent assessment approach in phases. Whereas the initial challenge consists of selection of an adequate patient population, subsequently all determinants of disease (signs, symptoms, comorbidity, and other diagnostic factors) and factors influencing the decision to use a test (patient burden and cost) are considered.**

MANY NEW DIAGNOSTIC tests have been developed and introduced into clinical practice in recent years. Consequently, a need has developed for proper assessment of tests before implementation. While consensus exists as to necessity of a phased assessment of a new drug, this is not the case for diagnostic tests. Three factors are of central importance in diagnostic test assessment: (1) the selection of the proper study population, (2) the determination of diagnostic power with its related statistical analysis, and (3) the comparison of the new test with current diagnostic tools.

For assessment of a test, usually the test results of healthy people are compared with those of patients already known to have a given disease: the easily accessible population. In clinical practice, however, the test is used to distinguish between the presence and absence of disease among patients having certain symptoms and manifesting particular signs. In this so-called "indicated" population, it is more difficult to discriminate between suspected patients with the illness and those without, because it is highly probable that the test outcome is associated with complaints, signs, or symptoms of the study patients. It is the assessment in an indicated population in which we are ultimately interested.[1]

The second point to be stressed in assessment studies is the definition of diagnostic power and the associated statistical and analytical methods. Commonly, diagnostic power is ascribed to a qualitative diagnostic test if it yields a higher test result more frequently in the diseased group than in the nondiseased group. The results of quantitative tests often are dichotomized, usually at the 95th percentile of the test result distribution in healthy persons,[2] so as to make the test a qualitative one. Dichotomization deprives a test of its full informative content, and thus may lead to substantial loss of information. Furthermore, it makes assessment of results dependent on the cutoff point and susceptible to selection bias.[3] Selection bias can occur as a result of differential verification for patients with positive and negative test results,[4] or as a result of studying only a part of the entire disease spectrum,[5] as is done in an easily accessible population.

Finally, it also is important to determine the relative value of the new test and existing diagnostic tools to de-

From the *Department of Medical Informatics and Epidemiology, University of Nijmegen, and the †Department of Diagnostic Radiology, University Hospital Nijmegen, Nijmegen, The Netherlands.
Reprint requests: Yvonne T. van der Schouw, PhD, Department of Epidemiology, Utrecht University, PO Box 80035, 3508 TA Utrecht, The Netherlands.
Received March 7, 1995, and accepted for publication, after revision, May 12, 1995.

termine if the new test has diagnostic valve. It might be possible to eliminate one or more of the older tests or, indeed, the new test. In this article, we elaborate on these factors and incorporate them into guidelines for a phased assessment of new diagnostic tests.

Diagnostic Process

Generally, patients visit a doctor because they are experiencing certain symptoms. The doctor will, implicitly, list diseases that are typically attended by the complaints, signs, and symptoms that the patient is encountering.[6] A probability of disease presence, depending on the frequency with which the doctor sees these diseases, is assigned to each disease on the list. If one probability is high enough, treatment will be initiated. If no probability is high enough, a diagnostic test will be performed. Whether a probability is sufficiently high to warrant treatment or the use of a diagnostic test cannot be determined in general terms. It highly depends on the benefit of the treatment, the risk or cost of the diagnostic test, and the cost of treatment for patients who do not actually have the disease, but may be treated because of a false-positive test result.[5,7] The assigned probabilities of disease presence are adjusted for each disease on the basis of a test result.

Current Assessment Practice

For assessment of a test, the test results of healthy people are compared with those of patients already known to have the disease. Diagnostic power is ascribed to a qualitative diagnostic test if it yields a higher test result more frequently in the diseased group than in the non-diseased group. For example, mean serum concentrations of carcinoembryonic antigen were higher in patients with colorectal carcinoma than in healthy individuals, therefore it was long believed that carcinoembryonic antigen might be a suitable diagnostic test for colorectal carcinoma. The results of quantitative tests often are dichotomized, usually at the 95th percentile of the test result distribution in healthy persons,[2] so as to make the test qualitative. Frequently, assessment is considered appropriate when the sensitivity and specificity of a new test are determined.

Problems With Current Assessment Practice

The well-known parameters, sensitivity and specificity, of the applied test, to be calculated from a basic table such as Table 1, are useful for updating the probability of disease presence according to Bayes' theorem.[5,8] Table 2 shows how the probability of disease in the presence of a positive test result is calculated with Bayes' theorem in the notation of conditional probabilities. Thus, in diagnosis, a strong need exists to know the information a test result yields (eg, by giving the sensitivity and specificity).

TABLE 1. Information of a Dichotomous Diagnostic Test as Presented in a Fourfold Table

		Disease		
		Present	Absent	
Test result	Positive	320	60	380
	Negative	80	540	620
Total		400	600	1000

Sensitivity: proportion of diseased with positive test result = 320/400 = 80%.

Specificity: proportion of non-diseased with negative test result = 540/600 = 90%.

Prevalence of disease in test-positive group = 320/380 = 84%; often referred to as positive predicted value.

Prevalence of disease in test-negative group = 80/620 = 13%; also known as 1 minus negative predictive value.

In assessment studies of new tests, estimates of these test "characteristics" predominantly are presented as assessment results. Unfortunately, sensitivity and specificity do not always provide the relevant information for this purpose, mainly because of their dependency on cutoff points and their susceptibility to selection bias.

Many diagnostic tests yield results that are not just positive or negative, but have a wider range, such as the percentage occlusion of an artery as judged from angiography. To calculate sensitivity and specificity, the possible test results must be dichotomized. A cutoff point for test positivity must be chosen; for example, a result higher than 70% is positive and should result in patient referral for surgery. By doing so, the different diagnostic information of 50% stenosis or 90% stenosis is not taken into account. Thus, dichotomization deprives a test of its full informative content and may lead to substantial loss of information. Furthermore, it makes assessment results dependent on the cutoff point and susceptible to selection bias.[4]

Selection bias can occur as a result of differential verification for patients with positive and negative test results,[4] or as a result of studying only a part of the entire disease spectrum.[5] Results of a new test usually are compared with results from patients already diagnosed with the disease: the easily accessible population. In clinical practice, however, the test is used to distinguish between the presence and absence of disease among patients having certain symptoms and manifesting particular signs. In this indicated population, it is more difficult to discriminate between suspected patients with the illness and those without because test results often are associated with the complaints, symptoms, or signs of study patients. Consequently, sensitivity and specificity results of tests that are applied to a population that was selected differently than the population in which the test was assessed cannot be extrapolated to the new population.[9] It

TABLE 2. Bayes' Theorem

Bayes' theorem in terms of conditional probabilities	Bayes' theorem in a more practical notation
$P(D^+ \mid T^+) = \frac{P(D^+) \times P(T^+ \mid D^+)}{P(D^+) \times P(T^+ \mid D^+) + P(D^-) \times P(T^+ \mid D^-)}$	$P(D^+ \mid T^+) = \frac{P(D^+) \times Sensitivity}{P(D^+) \times Sensitivity + 1 - P(D^+) \times Specificity}$

D: disease outcome (present/absent); T: test result (positive/negative).

is the assessment result from an indicated population in which we are ultimately interested.[10,11]

Consider, for example, a population in which 60% has complaints. The results of the diagnostic test used correlate perfectly with the presence or absence of complaints. Two thirds of patients with complaints have the disease, and three quarters of patients without complaints do not have the disease. The prevalence of disease is 50% (Table 3). In this population, the sensitivity rate of the test is 80%, and the specificity rate is 60%. In a second population, only 40% has complaints, therefore, the distribution of disease presence and absence in patients with and without complaints is the same as in the first population. Thus, two thirds of the patients with complaints have the disease and three quarters of the patients without complaints do not have the disease (Table 4). In this population, the sensitivity rate of the same test is 64%, and the specificity rate is 78%. Thus, the composition of the patient population is one of the determinants of sensitivity and specificity. These parameters are not constant characteristics of a test, but vary with the distribution of complaints, symptoms, and signs.

In our opinion, it also is important to establish the relative value of the new test and existing diagnostic tools to determine whether the new test contributes as a diagnostic tool. It might be possible to eliminate one or more older diagnostic tests or, indeed, the new test.

When assessing the diagnostic value of a new test, it must be compared with a generally accepted reference method, or a gold standard. In many instances, gold standards do not exist or are not ethically justifiable to carry out on all patients (eg, surgery or autopsy). In such instances, appropriate fallible reference methods may be used, such as the generally accepted diagnostic reference test, clinical follow-up for a fixed time period, or response to therapy.

Assessment in Phases

The main challenge in diagnostic test assessment is selection of an adequate study population. Ultimately, this population consists of indicated patients. To avoid waste of effort and money, it must be established as soon as possible whether a diagnostic test is potentially worthwhile. This can be achieved by a prudent assessment in several phases.[12–17] It is not necessary that all new tests undergo the entire assessment procedure. In some instances, it is established quickly that test outcomes among diseased and nondiseased patients are virtually identical. In that case, the test is not yet adequate.

Test Development

Strictly speaking, the development of a diagnostic test does not fall within the scope of test evaluation and will not be elaborated further. Nevertheless, it is important that some facts of the new test are known before actual assessment starts. Among these are the following.

1. Minimum detection level and cross-reactivity in the case of a biochemical test: for a useful diagnostic test, all individuals undergoing the test, regardless of whether they have the disease, should have test results that are higher than the minimum detection level of the test. Furthermore, the biochemical test should only react with the substance of interest (eg, a specific serum tumor marker) and not with other substances.
2. Measurement errors: large random measurement errors make potential diagnostic tests useless beforehand.
3. Repeatability, which must be high, otherwise the test will never meet diagnostic assessment standards.

TABLE 3. Hypothetical Distribution of Test Results and Disease Presence-1

		Disease present	Disease absent	Total
Test result	Positive	40	20	60
	Negative	10	30	40
Total		50	50	100

TABLE 4. Hypothetical Distribution of Test Results and Disease Presence-2

		Disease		
		Present	Absent	Total
Test result	Positive	27	13	40
	Negative	15	45	60
Total		42	58	100

4. Required personnel and equipment.
5. Acceptability of the test for patients. A test can have very good discriminative power; however, if it poses risk of morbidity or even mortality to the patients, such as coronary angiography, application of the test will be limited only to patients with a very strong suggestion of disease presence.
6. Dose and pharmacokinetics in the case of a pharmaceutical test, etc.

Test Application in an Easily Accessible Population

To carry out an assessment in an easily accessible population can only be useful from a pragmatic and logistic point of view. If the test performs badly here, which is not uncommon, it will certainly perform worse in the indicated population. This phase can be regarded as a "quick and dirty" assessment method. An easily accessible population could be, for instance, a population of diagnosed patients and healthy individuals. We are not concerned with the occurrence of selection bias or other problems at this point. If a test does not have sufficient discriminative power in this phase of diagnostic test assessment, the assessment procedure can be stopped. The test needs further development.

Test Application in the Indicated Population

This essential assessment phase of the "indicated" population should take place in a population of patients who are judged eligible for the new test in a real clinical situation because of the suspicion that they have the illness at issue. In this phase, therefore, we are concerned with the selection of the relevant patient population. Adequate selection of patients is achieved by gathering a consecutive series of all as yet undiagnosed patients suspected of having the disease at issue on the ground of their complaints, signs, and symptoms.[18]

Diagnostic Profile

At this stage of the assessment process, it is important to establish the contribution of the new test to the existing diagnostic arsenal, including other diagnostic tests, the nonclinical profile (eg, age, gender), the clinical profile (complaints, symptoms, and signs) and comorbidity. If the new test cannot add diagnostic power to the existing arsenal, it should be questioned seriously whether introduction of the test into clinical practice is desirable. Only if the test has other relevant characteristics, such as cheapness or safeness, introduction can be considered. In this phase, it also is necessary to study only patients of the proper indicated population. To obtain the diagnostic information of the new test alone, the outcomes of all other diagnostic tests in use must be known as well, which necessitates the concurrent measurement of those tests.

Medical Technology Assessment

The last challenge for the new test is a medical technology assessment, including selection and order of diagnostic tests, cost-effectiveness, patient's utilities or preference measurements, etc. The new test can be used as the first test after the history has been taken and physical examination has been carried out. However, it is also possible that the new test is so expensive that to apply it to all patients is not justified, such as in cases where the natural course of the disease is not life-threatening. In this phase, questions of therapeutic efficacy (did treatment or patient management change?), patient outcome efficacy (did life expectancy improve?) and, if possible, societal efficacy (are costs acceptable for society?) should be answered.[19]

Because there are so many questions involved in the final phase of test assessment, the best way to deal with them probably is to carry out a clinical trial in the indicated population, with random allocation of the new test. Test results will have therapeutic consequences and the only proper way to study such an intended effect is a randomized controlled trial.

Necessary Knowledge

Before test assessment can be carried out in any study population, certain information has to be made explicit. First, the population in which the test is assessed should be defined carefully. Also, the ultimate realization of the study population has to be described, while indicating whether the investigators succeeded in gathering all consecutive patients or not. This gives an indication of the potential selection bias contracted by the actual study population. Even an indicated population will yield biased assessment results when selective uptake or loss of participants has taken place.

Secondly, the diagnostic test must be described clearly and test results presented accordingly. This includes specifying the type of results it yields. Test results may be measured qualitatively or quantitatively. Many tests are quantitative but have been converted to qualitative tests. An example of such a test is the hemoglucotest, which measures glucose concentration in blood. The test is designed so that at a certain blood glucose concentration the color of the test strip changes to blue.

If tests are inherently qualitative, it is possible to assign a judgment probability of the presence of disease to the test result to create a quantitative test result. In the Dutch screening program for breast cancer with mammography it is common practice to judge the presence of malignant lesions on mammograms in terms of "not present, not probable, indecisive, probable, certain."

If judgment of the test result is done by different observers, they should be standardized to diminish interobserver variation as much as possible. New observers

should be trained by experienced observers and every once in a while consensus meetings might be arranged. However, interobserver variation will never be eliminated entirely. If an apparatus is simple to read, interobserver variation may not be a problem, but systematic measurement errors of the apparatus should be known for the machine to be calibrated as often as necessary. Standardization is important to eliminate systematic differences, but for tests involving a reader's subjective interpretation, interobserver variation cannot be eliminated. Here, consensus and dual reading are tools for standardization.

Finally, the manner of verification of disease status should be stated for all patients in the investigation. Sometimes all patients can be subjected to the gold standard, but often, for ethical reasons, clinical follow-up for a restricted period of time must be used as the criterion, at least for part of the study population, and in particular for patients with negative test results. Occasionally, verification relies on the prescription of and subsequent reaction to a medication, for instance a certain pharmaceutical. Calculation of test parameters such as sensitivity and specificity should not be restricted to those patients who were subjected to the real gold standard. In that case, the parameter estimates will be biased; this type of bias is called verification bias.[20]

All relevant patient characteristics, test results, and the ultimate diagnosis should be registered individually for every patient, including missing data. With all this information available, insight is gained into the potential bias of assessment measures, and the extent and direction of possible bias.

Statistical Analyses

For qualitative tests—measuring the test result on a presence/absence scale—commonly used measures of test performance are sensitivity and specificity, calculated from a fourfold table, or likelihood ratios, to be calculated from a fourfold table or a 2 × k table—in case the test results are measured in k categories (Table 1).[21] For quantitative tests, with test results in more than two ordinal categories or on a continuous scale, it is recommended to present all the information they provide. This can be achieved by presenting receiver operating characteristic (ROC) curves[22] and cumulative probability distributions.[23]

TABLE 5. Distribution of Duplex and Angiography Results of Carotid Arteries in a 2*k Table

		Angiography % stenosis		
		71–99	≤70	Total
Duplex % stenosis	71–99	7	22	29
	51–70	10	1	11
	31–50	0	15	15
	≤30	0	22	22
Total		17	60	77

TABLE 6. Distribution of Duplex and Angiography Results of Carotid Arteries in a 2*2 Table

		Angiography % stenosis		
		71–99	≤70	Total
Duplex % stenosis	>30	17	38	55
	≤30	0	22	22
Total		17	60	77

An ROC curve displays sensitivity versus 1 minus specificity at as many cutoff points as possible.[22,24] Receiver operating characteristic curves are thus independent of cutoff points and yield the area under the curve as a summary measure of diagnostic performance.[25,26] The construction of ROC curves is as follows. Consider the Duplex test for assessment of carotid stenosis. This test was assessed in 77 carotid arteries end carotid angiography was performed in all of them as the gold standard. The new test must be able to determine whether a patient should have a carotid endarterectomy (stenosis 71 to 99%) or not (Table 5). At each possible cutoff point the table is condensed to a 2 × 2 table and sensitivity and specificity are calculated. For example, when the Duplex criterion of >30% stenosis is used as cutoff point for test positivity, the table simplifies to Table 6. The sensitivity then is 100% and the specificity is 37%. Next, sensitivity and 100% minus specificity are plotted against each other in an ROC curve (Figure 1).

Furthermore, different tests assessed in one population can be compared easily, and the ROC curve is relatively stable to verification bias.[3,4] Cumulative frequency distributions provide clinicians with the sensitivity and specificity of a test at any cutoff point requested.

There is a tendency to present the prevalence of illness by groups of test results,[1] because this is the proportion we are naturally interested in: given a certain test result, what is the probability that disease is present? This makes sense only when the actual study population has been gathered in ignorance, and therefore independently, of the true disease status.

Prevalence of disease can be calculated for patients with a positive test result and for patients with a negative test result, as well as for the various categories of test results for a quantitative test. In the case of a perfectly discriminating test, prevalence in the lower categories of the test result would be 0, whereas the prevalence of disease

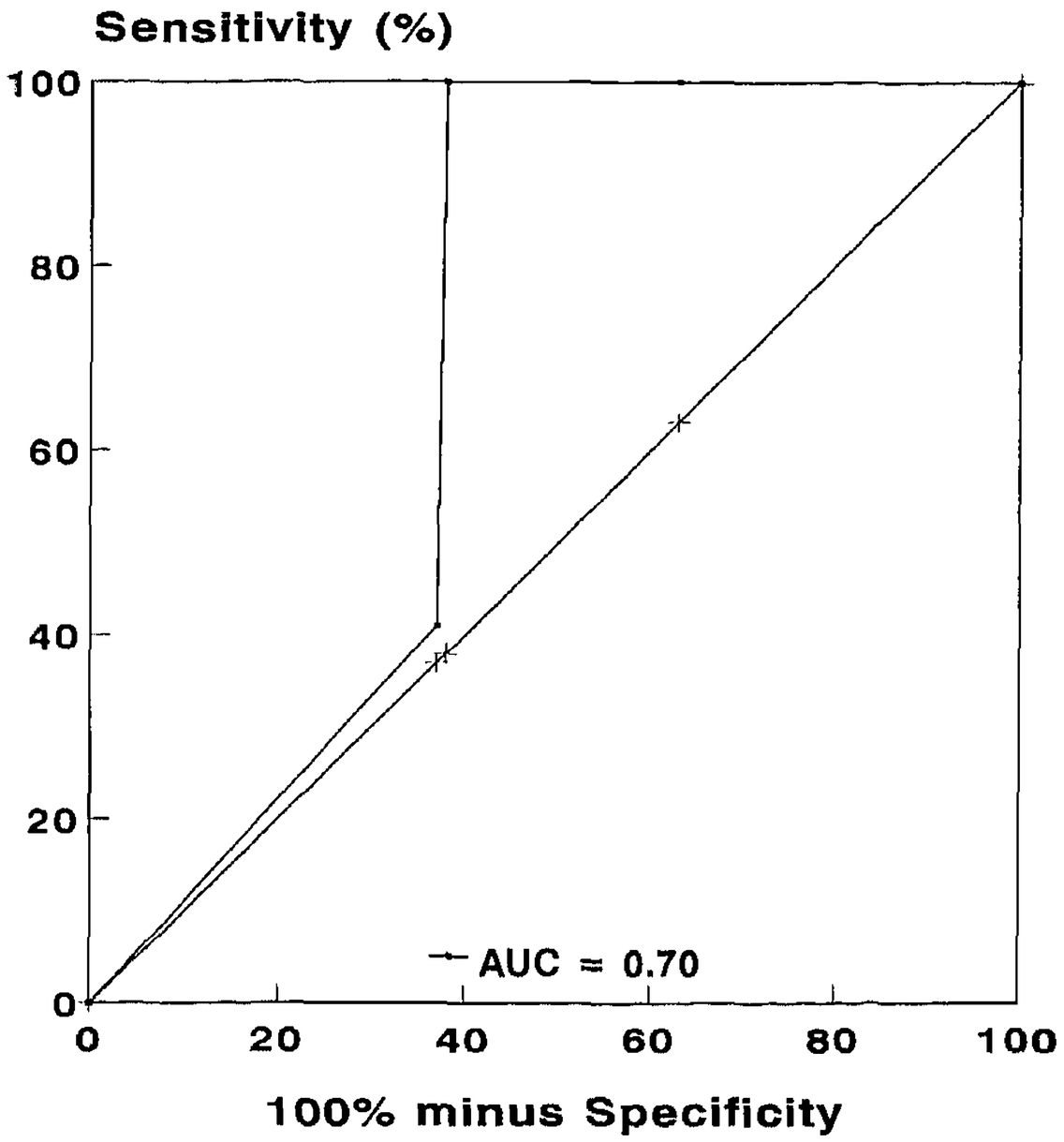

Fig. 1. Receiver operating characteristic curve for Duplex measurements of carotid artery stenosis.

in the higher would be 1. If the prevalence of disease conditional on clinical and nonclinical profile, comorbidity, and other diagnostic test results is required, logistic regression analysis is an attractive way of estimating this so-called prevalence function.[10] According to the presence or absence of a feature or the result of a diagnostic test for an individual patient, data is substituted in the prevalence function, and the probability of having the disease can be calculated.[27]

As in all analyses, missing values on test outcome or disease status have to be dealt with properly, for instance as separate categories, rather than removed from the analyses and wholly disregarded.[28]

Conclusion

Guidelines have been presented with recommendations for the design and statistical analysis of assessment studies. With respect to the patient population, it was explained that patients with a clinical indication for the test form the only adequate population. Patients have an indication for a test because they are suspected of having the disease that the test is supposed to diagnose. With respect to the statistical analysis, a plea for the use of ROC curves and logistic regression analysis was made. Receiver operating characteristic curves use the full information content of a diagnostic test, in case the test is measured on an ordinal or continuous scale. With logistic regression analysis, the contribution of the new test to the existing diagnostic arsenal can be estimated.

Good guidelines for diagnostic test assessment could prevent the introduction of disappointing diagnostic tests before they are fully implemented in clinical practice.[29] Although a phased assessment is not foolproof, it is highly likely that the chances of detecting a useless test are increased by introducing such an assessment procedure. This is illustrated aptly by the example of phased medical drug evaluation.

Acknowledgments

The authors thank Prof. O. S. Miettinen and Dr. J. J. Caro for their helpful discussions, and Drs. F. Joosten, D. van der Heijde, and R. Laan for help on earlier versions of this manuscript.

References

1. Miettinen OS. Theory of medicine: At the core of post-Flexnerian education in medicine? Amsterdam, The Netherlands: Free University; 1987. Inaugural address.
2. Linnet K. Comparison of quantitative diagnostic tests: Type I error, power, and sample size. Stat Med 1987;6:147–158.
3. Schouw van der YT, Straatman H, Verbeek ALM. ROC curves and their AUC of dichotomized tests: Empirical findings for logistically and normally distributed diagnostic tests. Med Decis Making 1994; 14:374–381.
4. Diamond GA. ROC steady: A receiver operating characteristic curve that is invariant relative to selection bias. Med Decis Making 1987;7:238–243.
5. Sackett DL, Haynes RB, Tugwell P. Clinical epidemiology: A basic science for clinical medicine. Boston, MA: Little, Brown, and Co; 1985.
6. Wulff HR. Rational diagnosis and treatment. Oxford, England: Blackwell Scientific Publications; 1976.
7. Doubilet PM. Statistical techniques for medical decision making: Applications to diagnostic radiology. AJR Am J Roentgenol 1988; 150:745–750.
8. Sox Jr HC, Blatt MA, Higgins MC, Marton KI. Medical Decision Making. Boston, MA: Butterworths; 1988.
9. Hlatky MA, Pryor DB, Harrell FE, Califf RM, Mark DB, Rosati RA. Factors affecting sensitivity and specificity of exercise electrocardiography. Am J Med 1984;77:64–71.
10. Miettinen OS. Theoretical epidemiology: Principles of occurrence research in medicine. New York, NY: John Wiley & Sons; 1985.
11. Gur D, King JL, Rockette HE, Britton CA, Thaette FL, Hoy RJ. Practical issues of experimental ROC analysis: Selection of controls. Invest Radiol 1990;25:583–586.
12. Guyatt G, Drummond M. Guidelines for the clinical and economic assessment of health technologies: The case of magnetic resonance. Int J Technol Assess Health Care 1985; 1:551–556.
13. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of medical technologies. Can Med Assoc J 1986; 134:587–594.
14. Freedman LS. Evaluating and comparing imaging techniques: A review and classification of study designs. Br J Radiol 1987; 60: 1071–1081.
15. Nierenberg AA, Feinstein AR. How to evaluate a diagnostic marker test: Lessons from the rise and fall of dexamethasone suppression test. JAMA 1988;259:1699–1702.
16. Begg CB. Experimental design of medical imaging trials: Issues and options. Invest Radiol 1989;24:934–936.
17. Köbberling J, Trampisch HJ, Windeler J. Memorandum for the evaluation of diagnostic measures. J Clin Chem Clin Biochem 1990;28:873–879.
18. Schouw van der YT, Dijk van R, Verbeek ALM. Problems in selecting the adequate patient population from data files for assessment of new diagnostic tests. J Clin Epidemiol 1995;48: 417–422.

19. Thornbury JR, Fryback DG. Technology assessment: An American view. Eur J Radiol 1992;14:147–156.
20. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978;22:926–930.
21. Rifkin RD, Hood WB. Bayesian analysis of electrocardiographic exercise stress testing. N Engl J Med 1977;297:681–686.
22. Weinstein MC, Fineberg HV. Clinical decision analysis. Philadelphia, PA: WB Saunders Co; 1980.
23. Zwetsloot-Schonk JHM, Hermans J, Frohlich M, Snitker P, Soeverijn JHM, Zwartendijk J. Sensitivity and specificity of acid phosphatase to detect prostate cancer using data from a hospital information system. Methods Inf Med 1990;29:213–219.
24. Hanley JA, McNeil BJ. The meaning and use of the area under the receiving operating characteristic (ROC) curve. Radiology 1982;143:29–36.
25. McNeil BJ, Hanley JA. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. Med Decis Making 1984;4:137–150.
26. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. Invest Radiol 1989;24:234–245.
27. Schouw van der YT, Velden van der MTW, Hitge-Boetes C, Verbeek ALM. Diagnosis of hypertrophic pyloric stenosis: Value of sonography when used in conjunction with clinical findings and laboratory data. AJR Am J Roentgenol 1994;163:905–909.
28. Hosmer Jr DW, Lemeshow S. Applied logistic regression. New York, NY: John Wiley & Sons; 1989.
29. Schouw van der YT, Segers MFG, Smits L, Thomas CMG, Verbeek ALM, Wobbes TH. Towards a more standardized assessment of diagnostic tumour markers. Int J Oncol 1993;3:979–985.